TECHN 28

Publication technique de la SmalS-MvM

09/2004

La préservation à long terme de l'information numérique

1. Introduction

Isabelle Boydens
est consultante à la section Recherches. Docteur en Philosophie et Lettres, orientation « Sciences de l'information et de la documentation», elle enseigne à l'Université Libre de Bruxelles. Ses travaux portent sur l'analyse de la qualité des bases de données, les systèmes de méta-information (tels que les « glossaires électroniques » développés dans le cadre du projet DMFA - DRS), les applications documentaires, les méthodes d'indexation et de recherche et les systèmes collaboratifs.

Contact: 02/509.59.92

isabelle.boydens@smals-mvm.be

1.1. Une problématique séculaire

La question de la conservation à long terme de l'information numérique[1] est devenue cruciale dans le monde des entreprises et des administrations. Nous proposons ici d'approfondir les facteurs à l'origine de la détérioration de l'information numérique ainsi que les stratégies destinées à y remédier. Nous verrons que cette question implique la prise en compte de nombreux facteurs d'ordre « hardware », « software » ou conceptuel[2].

Notons, au seuil de cette étude, que la problématique n'est pas neuve. En effet, au début du Moyen-âge, le passage du papyrus au parchemin a permis une consolidation des supports physiques[3]. Mais le recours massif aux peaux de bêtes soulevait des difficultés économiques. C'est ainsi que sont apparus les palimpsestes, parchemins dont on effaçait l'écriture antérieure pour écrire un nouveau texte, à l'instar des procédures qu'offrent actuellement nos disques optiques réinscriptibles. Certains écrits ont donc été effacés. Avant l'invention de l'imprimerie, les systèmes d'information se déployaient avec les générations de « moines copistes » qui recopiaient les manuscrits de siècle en siècle. Ces manuscrits ne nous sont souvent parvenus que sous forme partielle : il arrive que l'original soit perdu et que l'on ne dispose que d'un ensemble incomplet de copies divergentes (suite à des erreurs, volontaires ou non, commises par les copistes : passages transformés, omis ou ajoutés). Afin de reconstruire un texte qui soit le plus proche possible de l'original, l'historien construit un « *stemma codicum* » (ou généalogie des données), technique empruntée à la philologie. Dans tous les cas, la reconstruction est conjecturale. Autre exemple, plus récent : en 2000, un programme fut mis en oeuvre sous l'égide de la Bibliothèque Nationale de Berlin en vue de restaurer les partitions originales de l'œuvre de Jean-Sébastien Bach. Le compositeur utilisait en effet une encre ferrée qui libère, en s'oxydant, un acide très agressif pour le

[1] Nous considérons ici le terme numérique comme un synonyme du terme anglais "digital".

[2] Ce techno est une version remaniée d'un article sur le même thème : Boydens I., La conservation numérique des données de gestion. *Revue Document Numérique* (Numéro spécial : « *Archivage et pérennisation* »), Paris : Editions Hermès Science, (paru en septembre 2004).

[3] Chartier R. et Martin H.-J., éds, *Histoire de l'édition française. Tome I. Le livre conquérant. Du Moyen-âge au milieu du XVIIème siècle,* Paris, Fayard-Cercle de la Librairie, 1989.

papier et rend progressivement certaines partitions illisibles[4]. Toutefois, il faut bien reconnaître qu'on a pu conserver des informations sur support « papier » pendant des siècles, alors que la durée de vie de l'information numérique, si l'on n'y prend garde, peut s'avérer beaucoup plus courte.

1.2. La conservation des données numériques

La problématique est plus complexe s'agissant des données numériques. En effet, les séquences de bits dont elles sont composées ne sont pas « auto-explicatives », contrairement à un texte figurant sur un support papier ou mural auquel l'utilisateur a un accès « direct » (s'il en comprend la langue ; souvenons nous que les hiéroglyphes n'ont été déchiffrés qu'au début du 19ème siècle par Champollion). L'information numérique, qu'elle soit « *born digital* »[5] ou qu'elle résulte d'une opération de numérisation[6] est fragile. Citons un exemple : le Département de la Défense américain a dû mettre en place une coûteuse opération de « migration » en vue de restaurer les fichiers relatifs aux coordonnées géographiques des bombes lâchées durant la guerre du Vietnam. Ces données, stockées à l'époque dans une base de données « propriétaire », étaient devenues illisibles : quelques années plus tard, plus aucun logiciel du marché ne permettait de les traiter sans générer d'incohérences. L'enjeu était considérable puisqu'il s'agissait de déminer et d'identifier les bombes qui, n'ayant pas explosé pendant ou après la guerre, pourraient occasionner de nouvelles victimes. La correction des incohérences fut possible grâce aux services des « *National Archives* » qui avaient conservé le format propriétaire original et l'avaient périodiquement transféré dans des environnements plus modernes, tant sur le plan logiciel qu'au niveau des formats[7]. L'information numérique repose ainsi sur une complexe « *chaîne de médiation à la fois matérielle et logicielle* »[8].

Dans le point 2 qui suit, nous définissons deux composantes importantes de la problématique : quelle information veut-on conserver et pour combien de temps ? Le point 3 présente les facteurs à l'origine d'une détérioration de l'information numérique (obsolescence du hardware, des supports physiques, des composantes logicielles et des formats, ces facteurs interagissant). Nous évaluons ensuite (point 4) les stratégies de conservation proposées à ce jour (« refreshing », migration, préservation technologique, émulation, encapsulation et recours aux méta-données). Nous ajoutons à ces techniques, traditionnellement évoquées, un autre point : dans le contexte de la gestion d'applications administratives « vivantes »[9], nous indiquerons quelques pistes complémentaires applicables dès la création des données, en vue de faciliter leur conservation ultérieure, une fois celles-ci archivées, au terme de leur « cycle de vie ».

En conclusion (point 5), nous synthétisons la problématique et suggérons plusieurs recommandations méthodologiques adaptées au contexte de la sécurité sociale et, plus largement, de l'administration fédérale.

---

[4] Le projet, alors estimé à une durée de trois ans, fut doté d'un budget d'un million d'euros. Une technique de dédoublement des documents a été mise en œuvre : « *pour cela la partition est pressée entre deux feuilles enduites de gélatine, qui prennent grain à grain l'empreinte du manuscrit. Les « négatifs » sont ensuite fixés sur une nouvelle page* ». *Programme pour sauver les partitions de Bach de l'autodestruction,* AFP (dépêche), 25/02/2000. Voir aussi : http://www.rtbf.be/matieregrise/emissions/mg23/textes/ailleurs.html (consulté le 18 mars 2004).

[5] C'est-à-dire directement produite sous forme électronique, via un traitement de texte ou un système de gestion de base de données, par exemple.

[6] Un document papier étant par exemple converti en fichier numérique suite à une opération de reconnaissance optique des caractères.

[7] Ruggiero A. (éd.), « Preservation of Digital Memory. Risks and Emergencies. Six Case Studies », *The Future of Digital Memory and Cultural Heritage*, Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche, Florence, 16 et 17 octobre 2003, p. 35-36.

[8] Masanes J., « L'information technique nécessaire à la préservation à long terme des documents numériques », *International Preservation News, Newsletter of the IFLA Core Activity on Preservation and Conservation*, n°29, mai 2003, p. 12.

[9] Nous entendons par "vivantes" des données qui, n'étant pas arrivée au terme de leur cycle de vie, sont toujours exploitées à des fins de gestion opérationnelle.

2. Objet et durée de la conservation

Deux questions se posent d'emblée. Que veut-on conserver ? Et pour combien de temps ?

L'archivage des données et donc, leur conservation, relève du « *record management* ». Le terme anglo-saxon « *record* » englobe tout type de document « *produced or received by a person or organisation in the course of business and retained by that person or organisation* »[10], quel que soit le format, le support ou la nature plus ou moins structurée de l'information[11]. La norme européenne en la matière ajoute : « *A key feature of a record is that it cannot be changed* ». Par ailleurs, les « *records* » revêtent généralement une valeur légale[12]. Dans ce contexte, l'objet de la conservation se limite aux documents, arrivés au terme de leur « cycle de vie », et dont le contenu est figé et ne fera plus l'objet de modifications.

Dans cet article, nous étendons le champ aux données de gestion « vivantes », susceptibles de modification car, nous le verrons, la problématique de la conservation à long terme de l'information numérique prend naissance en T_0, lors de la création de l'information.
Nous proposons d'évoquer les stratégies de conservation selon l'état d'une donnée dans son cycle de vie. Parmi les informations dont la conservation est jugée utile, on peut distinguer le « temps court » des données de gestion, le « temps intermédiaire » de la préservation légale et le « temps long » de la conservation historique :

- Les données « de gestion » administratives s'inscrivent dans un « temps court » durant lequel les « *records* » se constituent : par exemple, lorsqu'une lettre officielle est générée suite à l'extraction d'informations issues d'une base de données et au terme d'une phase de validation par « *workflow* ». A ce stade, les documents ont un impact opérationnel direct sur le réel observé (les citoyens assujettis à l'administration fédérale, par exemple). Quelle est la durée de ce temps court ? Combien de temps dure le présent ? Le temps que les « données vivantes » fassent leur office « de gestion » : dans le domaine administratif, ce temps est théoriquement inférieur à la période de prescription[13] (à titre d'exemple : cinq ans dans le secteur de la sécurité sociale belge).
- Ensuite, on distingue un « temps intermédiaire », couvrant la période légale de prescription. Les « *records* » ne sont plus modifiés mais peuvent avoir un impact sur le réel et impliquer la génération de nouvelles données, en cas de procès et de réouverture d'un dossier. Ainsi, la récente affaire *Enron* aux Etats-Unis a montré l'importance de l'information d'entreprise : depuis 2003, la loi Sarbanes-Oxley réglemente aux USA la conservation des documents (incluant les e-mails) pour toute entreprise cotée en bourse[14].
- Enfin, se profile le « temps long » au cours duquel les « records » devraient être conservés « à perpétuité », non plus à des fins légales, mais en tant que source potentielle pour les historiens des temps futurs.

---

[10] *MoReq Specification, Model Requirement for the Management of Electronic Records,* IDA (Interchange Data between Administrations) Programme of the European Commission, mars 2001

(http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc, consulté le 18 mars 2004), p. 7-8.

[11] Que nous appellerons aussi « donnée » dans la suite de l'article.

[12] *Information and documentation – Records management*. Norme ISO 15489-1, 2001-09-05, 1ère édition, p. 9.

[13] La prescription est une disposition légale permettant de se libérer d'une obligation après écoulement d'une période déterminée. Le temps de prescription évolue lui-même dans le temps et est passé en 1996 de 3 à 5 ans dans le domaine de la sécurité sociale belge. Arrêté royal du 20 décembre 1996 modifiant l'arrêté royal du 23 novembre 1969 pris en exécution de la loi du 27 juin 1969 révisant l'arrêté-loi du 28 décembre 1944 concernant la sécurité sociale des travailleurs. *Moniteur belge*, 31 décembre 1996. La législation en la matière est parfois floue et variable d'un pays ou d'un secteur à l'autre. On trouvera des sites de références sur la question. Par exemple, pour les USA, celui de la National Archives and Records Administration (http://www.archives.gov/index.html) ou au Royaume-Uni, le site du Public Records Office (http://www.pro.gov.uk).

[14] Berdot V., Dossier « L'archivage d'e-mails ». Les éditeurs font feu de toutes lois, *O1 Informatique*, n°1756, 13 février 2004, p. 40.

Nous nous baserons sur cette structure temporelle en vue d'évaluer les différentes méthodes destinées à préserver l'information numérique. Nous verrons en effet que le recours à telles ou telles méthodes de préservation est plus ou moins pertinent en fonction de l'état d'un « *record* » dans son « cycle de vie ».

3. Les facteurs à l'origine d'une détérioration de l'information numérique

De nombreux éléments sont susceptibles d'altérer l'information numérique. Nous n'envisageons pas ici les catastrophes naturelles (incendies, inondations, tremblement de terre,...)[15] pour lesquelles des précautions bien connues, telles que la délocalisation des données, doivent être mises en place. Nous n'envisageons pas non plus les difficultés liées à l'interprétation de l'information[16]. Nous présentons successivement les points suivants : obsolescence du « hardware » et des supports physiques, obsolescence des composantes logicielles, fragilité ou hétérogénéité des formats.

3.1. Le « hardware » et les supports physiques

Les composantes matérielles (puces, circuits intégrés,...) d'un ordinateur se détériorent avec le temps. Par ailleurs, il se peut que tout ou partie du hardware d'un type d'ordinateur ne soit, à terme, plus supporté par la firme qui en assure la commercialisation.

A cela s'ajoute l'obsolescence des supports physiques et de leurs périphériques de lecture.

A propos des supports, citons le cas :

- des bandes magnétiques qui perdent progressivement leur charge magnétique (leur longévité estimée varie entre 10 et 30 ans) ;
- des CD-Worms qui s'usent avec la fréquence de lecture (certaines expériences affirment, sur la base de CD-Worms vieillis artificiellement, qu'ils dureraient cent ans) ;
- des microfilms dont la durée de vie, s'ils sont conservés de façon appropriée, dans un environnement sec et frais, peut s'étendre jusqu'à 500 ans. Cela dit, ce support particulièrement pérenne permet difficilement le traitement automatique des données : la conversion des média numériques vers ce medium analogique est peu pratiquée lorsque des fonctionnalités de recherche automatique sont requises pour accéder à un grand volume de données.

Les supports physiques sont en outre « *machine dependent* » ; ils requièrent un périphérique de lecture (associé à un « *driver* » logiciel). Or, ces éléments évoluent avec le marché. D'ores et déjà, on sait que certains formats de bandes magnétiques ne pourront plus être lus sous peu. A l'heure actuelle, certains « *notebooks* » ne supportent plus les « *floppy disks* » 3,5''. Combien de temps les lecteurs de CD-Rom vont-ils être maintenus quand on sait que les DVD avec une capacité de stockage au moins quatre fois plus importante sont considérés comme « le » medium de stockage « permanent » ? D'où cette « boutade » : « *If you are saving your data to CD-Rom, save a CD-Rom drive as well* »[17].

[15] En cas d'accident grave (la chute d'un avion sur un immeuble) touchant des centres financiers, par exemple, des sociétés spécialisées dans la restauration de documents interviennent juste après les pompiers et la police ; les documents endommagés par l'eau et le feu étant immédiatement congelés de façon à éviter une dégradation accrue des supports suite à l'évaporation. Perte d'information. Trouver la parade, *Archimag*, décembre 2003/janvier 2004, n°170, p. 26.

[16] Boydens I., « Les bases de données sont-elles solubles dans le temps ? », *La Recherche hors série* (« *Ordre et désordre* »), n°9, novembre-décembre 2002, p. 32-34.

[17] Davis M., « Memories are Precious », *Butler Opinion Wire*, 29 janvier 2004, p. 3.

3.2. Les composantes logicielles

Une application informatique ne peut être exploitée sans le logiciel qui l'a créée ou un logiciel analogue. L'obsolescence des composantes logicielles tient à ce qu'il n'y a pas systématiquement de compatibilité ascendante (« *backward compatibilty* »)[18] entre leurs versions successives. Avec les progrès technologiques, celles-ci se succèdent. Certains producteurs de logiciel fournissent des garanties quant à cette compatibilité ascendante, du moins pour un certain nombre de versions ; le passage d'une version à l'autre se faisant au cours d'une opération de migration, laquelle peut toutefois impliquer des pertes de données. Mais dans de nombreux cas, cette migration n'est pas prévue, ce qui implique de lourdes pertes d'informations (les liens entre les tables d'une base de données sont perdus, par exemple). Dans d'autres cas, certains logiciels « propriétaires » (parce que la firme qui les produit a fait faillite ou a fait l'objet d'un rachat) ne sont tout simplement plus suivis ni diffusés.

Dans la pratique, la problématique est plus complexe encore car il y a des interactions entre différentes couches logicielles et en particulier entre un logiciel applicatif et « *l'operating system* » (OS) se trouvant sur la machine. L'évolution des versions d'OS peut impliquer que tel ou tel logiciel n'est plus supporté dans toutes ses fonctionnalités. Par ailleurs il y a des interactions entre les évolutions des OS et celle des éléments matériels périphériques (une imprimante, par exemple), une incompatibilité pouvant apparaître entre une version d'OS et un périphérique donné. On trouve par ailleurs des interactions entre les versions logicielles et les types de supports, ainsi en ce qui concerne la lecture des bases Lotus Notes stockées sur CD-Rom pour laquelle une procédure spécifique doit être suivie[19].

Ajoutons que les pertes de données sont d'autant plus importantes que la structure de l'information est complexe : « *The more complex the digital resource, the greater the potential loss is likely to be. For example, interchanging the data held in geographical information system (GIS) databases and groupware databases could involve the loss of thousands of links that have taken years of effort to create and which represent the bulk of the value of the database*. »[20]. Idéalement, il faudrait disposer de standards internationaux afin de garantir l'homogénéité des interactions entre ces différentes composantes. Dans la pratique, ces évolutions sont souvent imprévisibles et tributaires des acteurs du marché.

3.3. Les formats et les types d'encodage

Les formats et les types d'encodage sont intimement liés aux aspects fonctionnels évoqués au point précédent. Afin de garantir la pérennité de l'information dont ils constituent la structure, les formats doivent être décrits par des normes publiques, documentées et idéalement certifiées par des organismes reconnus (Mercuri, 2003). Dans la pratique, deux types de difficultés se présentent :

---

[18] Cette compatibilité signifie qu'un logiciel d'une version v+1 permet de traiter une application créée avec la version v de ce même logiciel.

[19] « *Also, be aware that Notes databases on a CD can only be viewed by the same major version of Notes that the Notes database was indexed with. The view index and full text index are improved with each major version, i.e., R5 can't read R4 DB on CD, etc. To work around this with old CDs you may still want to view, you can copy the .nsf file to your local system, follow the procedure above, then copy all the files onto a new CDR because CDRW drives are so inexpensive now* ». http://www.keysolutions.com/NotesFAQ/howcd.html, consulté le 13 juillet 2004.

[20] Feeney M. (éd.), *Digital Culture : Maximising the Nation's Investment : a Synthesis of JISCO/NPO Studies on the Preservation of Electronic Materials*, Londres, National Preservation Office, 1999, p. 45.

- il arrive, d'une part, que des organismes dérogent à un standard, générant une hétérogénéité *de facto* ;
- d'autre part, certaines firmes informatiques diffusent d'emblée des formats propriétaires : l'utilisateur est alors tributaire de l'évolution de leurs versions et de la santé financière de la firme.

Ces phénomènes, assez prégnants au niveau du marché multimedia *(MoReq Specification,* 2001), génèrent une « volatilité » des formats et accroissent le risque d'obsolescence de l'information. Face à cette tendance, le recours au format XML du W3C offre des perspectives très intéressantes : la norme est fédératrice et présente l'avantage d'une séparation entre structure et contenu. Par ailleurs, le standard « de facto » et « non officiel » PDF s'est imposé à travers le monde en raison de son ouverture et de la gratuité du logiciel de lecture correspondant[21].

Les types d'encodage sont tout aussi fondamentaux. A cet égard, l'émergence d'Unicode et de la norme ISO/CEI 10646 pour le codage des caractères a permis d'élargir considérablement le spectre des langages pris en compte et, de là, de favoriser le traitement et l'échange homogène de l'information à travers le monde[22].

4. Les stratégies de conservation

Nous évoquons successivement les stratégies de conservation suivantes : « refreshing », migration, « technology preservation », « émulation et encapsulation », recours aux méta-données et, enfin, quelques principes de conception et de gestion. Ces stratégies sont complémentaires : aucune d'entre elles, considérée isolément, ne permettrait de résoudre l'ensemble de la problématique évoquée au point 3. Dans les conclusions (point 5), nous reprendrons dans un tableau synthétique la correspondance entre les différents facteurs à l'origine d'une détérioration de l'information numérique et la ou les stratégie(s) censées y remédier.

4.1. « Refreshing »

Le « *refreshing* » consiste à recopier l'information d'un support physique vers un autre, plus récent (par exemple, d'une bande magnétique vers une autre ou d'un support optique vers un autre). Cette approche permet de résoudre la question de l'obsolescence des supports physiques.

4.2. Migration

La migration consiste à convertir, via un programme, les données d'une configuration « *hardware/software* », en voie de devenir obsolète, vers une autre, plus récente. En raison de la multitude des interactions entre couches logicielles, la méthode peut toutefois entraîner des pertes d'information qu'il est parfois difficile d'identifier. Une opération de migration requiert dès lors des tests d'intégrité *a posteriori*, incluant un input intellectuel, en vue de vérifier la cohérence des données migrées par rapport à leur état antérieur. Largement pratiquée et complémentaire du « refreshing », la migration est l'option par défaut pratiquée au sein de nombreuses entreprises, administrations ou institutions en charge de la gestion de données « vivantes » ou archivées.

[21] Haut H., *Memo PDF*, Etude interne, SmalS-MvM, 14 septembre 1999.

[22] André J. et Hudrisier H., éds, « Unicode, écriture du monde ? », *Revue Document Numérique,* vol. 6, n°3-4/, 2002.

Cette procédure implique le suivi des modifications de versions logicielles dès la création des données traitées. Ce suivi inclut par ailleurs les interactions tolérées entre logiciels, *operating systems* et *drivers* de périphériques. Généralement, en ce qui concerne les « ténors du marché », des programmes de migration[23] sont fournis. Il est impératif de ne pas « sauter » de version majeure dans la mesure où la compatibilité ascendante n'est garantie que pour un nombre donné de versions. Dans tous les cas, l'opération peut être délicate comme le reconnaissent les producteurs de logiciels eux-mêmes : « *... but you also know that upgrading your databases and applications from currently installed Oracle products can be a complex and nerve-wracking job* »[24].

Dans une optique très constructive, certains encouragent la mise en place de services documentant les formats – faisant partie intégrante des éléments soumis à la migration - au fil de leurs versions [25]. La conservation sur support numérique étant problématique, on pourrait imaginer de conserver parallèlement cette documentation sur format papier. Il est utile par ailleurs de se référer aux organismes de certification en la matière (figure 1).

Name	Abbr.	URL	Coverage	Composition
American National Standards Institute	ANSI	www.ansi.org	Validates and promotes U.S. standards	Private, non-profit
British Standards Institute	BSi	www.bsi-global.com	British Standard 7799 for security policies and audits, also InfoSec certification	Business consortium
Institute of Electrical and Electronic Engineers Standards Association	IEEE-SA	standards.ieee.org	Over 140 security specific standards	IEEE membership
International Information Systems Security Certification Consortium	(ISC)²	www.isc2.org	Administers CISSP program for security professionals	Private, non-profit
International Organization for Standardization	ISO	www.iso.ch	International codes for standards management	National standards organizations
Internet Architecture Board	IAB	www.iab.org	Protocol standards for the Internet	IETF membership
National Institute of Standards and Technology	NIST	csrc.nist.gov	Computer security resources and standards for unclassified government data	U.S. government
National Security Agency	NSA	www.nsa.gov	Similar to NIST but for classified data	U.S. government
Organization for the Advancement of Structured Information Standards	OASIS	www.oasis-open.org	E-business standards	OASIS membership
Underwriter's Laboratories	UL	www.ul.com	Product safety testing and certification	Private, non-profit
World Wide Web Consortium	W3C	www.w3.org	Web-related interoperable technologies	W3C membership

Table. Computer security-related standards and certification organizations.

Figure 1. Quelques organismes de standardisation (source : Mercuri T. R., « Standards Insecurity », Communications of the ACM, décembre 2003, vol. 46, n°12, p. 11-19.).

[23] Une cause de migration peut être liée à l'accroissement de la taille autorisée des fichiers de données. On est dans certains cas passé d'un nombre maximum d'enregistrements par fichier de 16 millions à 2 milliards, ce qui a nécessité un accroissement de la taille des entrées d'index (répertoriant les clés primaires) : la migration est destinée à transformer les valeurs des anciennes entrées d'index en vue de les rendre compatibles avec les nouvelles. Dans d'autres cas, l'utilitaire de migration est directement intégré dans la nouvelle version et ses fonctionnalités ne sont pas explicites. En l'absence d'utilitaire et de compatibilité ascendante, il revient aux utilisateurs de découvrir les incohérences entre versions et d'y remédier dans la mesure du possible.

[24] Burke B., *Inside Oracle database 10g. The Great Migration Week experiment*, décembre 2003 (http://otn.oracle.com/pub/news/burke_10g_testing.html), consulté le 27/02/2004.

[25] Masanes J., *Op. cit.*, p. 16. Bien qu'incomplète, l'*Internet Assigned Numbers Authority* (IANA) va dans ce sens (http://www.iana.org/assignments/media-types/, consulté le 18 mars 2004)

4.3. « Technology preservation » ou musées d'ordinateurs

Certains proposent de conserver en un lieu toutes les formes d'*hardware* et de *software* utilisés à ce jour pour traiter les données[26]. L'approche n'est pas réaliste : la conservation en un seul lieu supprime la notion de réseau, cruciale en informatique. Ainsi les utilisateurs désirant lire leurs anciennes informations devraient voyager avec leurs données vers cette « Arche de Noé technologique ». Par ailleurs, les coûts en termes d'espace, de maintenance d'anciens équipements et de « *know how* » des techniciens seraient prohibitifs.

4.4. Emulation et encapsulation

L'émulation[27] repose sur l'encapsulation logique des données dans une couche « documentaire » décrivant l'environnement original (« *hardware/software* » et contextuel) qui a permis leur création. Le but est de recréer ensuite, sur cette base, l'environnement initial et de l'émuler (en imiter le comportement) sur une nouvelle génération de plateforme. Une expérience menée à la Bibliothèque Royale des Pays-Bas semble concluante mais les tests sont insuffisants à ce jour : « *l'émulation pose d'importants problèmes techniques (« simuler » une plate-forme informatique est loin d'être un problème trivial)* »[28]. La méthode est en effet considérée comme risquée et est critiquée pour son manque de généralité : « *Anyone relying solely on this strategy could be taking a significant risk. They would depending on the technical hability to the software engineers to emulate a specific environment and sustain it, and on the commercial viability of anyone providing such a service.* »[29]. « *Large-scale, long term emulation is not well-understood. It also has risks of loss of functionnality and other characteristics* »[30].

4.5. Recours aux méta-données

Plusieurs standards de méta-données[31] ont été proposés en vue de soutenir l'approche précédente (« émulation et encapsulation ») et, plus largement, de contribuer à la compréhension globale de l'environnement logiciel et matériel de l'information numérique. En tant qu'instrument d'interprétation de l'information, les méta-données en facilitent la conservation : elles permettent de documenter les opérations de « refreshing », ou de réaliser des migrations sur la base des caractéristiques (code logiciel, format, etc.) des données initiales à migrer. En ce sens, les méta-données sont utiles quelle que soit la méthode de conservation proposée.

---

[26] Kleinberg K. et Logan D., « Digital Preservation in Healthcare : Long-Term Accessibility ». *Gartner Research Note, Strategic Planning*, SPA-15-0907, 7 janvier 2002.

[27] Rothenberg J., An experiment in Using Emulation to Preserve Digital Publications, La Haye, Koninklijke Bibliotheek, 2000.

[28] Masanes J., *Op. cit.*, p. 13.

[29] Feeney M. (éd.), *Op. cit.*, p. 42.

[30] *MoReq Specification, Model Requirement for the Management of Electronic Records,* IDA (Interchange Data between Administrations) Programme of the European Commission, mars 2001 (http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc, consulté le 18 mars 2004), p. 85.

[31] Issu du grec "*meta*", qui signifie "*sur*", "*au dessus de*", le mot "méta-donnée" peut référer à une donnée documentant une autre donnée de plus bas niveau en vue d'en faciliter l'interprétation. Par exemple, à propos d'une donnée "date", une méta-donnée peut se présenter sous la forme d'une phrase en langage naturel spécifiant qu'il s'agit de la "date de facturation dans un système comptable".

Le format XML présentant l'avantage d'une séparation entre les valeurs et leur structure logique, quelle que soit la plateforme logicielle ou hardware, est de plus en plus préconisé à cette fin. Parmi les normes de méta-données, on trouve par exemple l'OAIS[32] (*Open Archive Information System*), développé sur l'initiative du CCSDS (*Comité Consultatif pour les Systèmes de Données spatiales*), standard ISO (ISO 14721 :2002). La norme propose, d'une part, des informations de représentation permettant de décrire le document numérique et son mode de production et, d'autre part, des informations décrivant les modes de préservation de l'objet numérisé[33].

Cela dit, si les méta-données sont indispensables[34], elles ne constituent pas « la » solution miracle pour deux raisons.

D'une part, les méta-données sont des données. Si les données numériques ne sont pas pérennes, pourquoi les méta-données le seraient-elles davantage ? Or, si les normes en matière de conservation numérique définissent la notion de donnée numérique (ou « record ») à conserver, la distinction entre les données et les méta-données est parfois esquivée : « *the distinction between data and its metadata can be unclear... these details of metadata usage are beyond the scope of the MoReq specification* »[35]. Certaines études considèrent implicitement que les méta-données revêtent une nature plus durable que les données dont elles sont censées faciliter la conservation. Or les méta-données n'échappent pas au foisonnement des standards : à côté de la norme OAIS, on trouve par exemple la norme METS (*Metadata Encoding and Transmission Standard)* : ces standards s'emboîtent-ils comme des « poupées russes »[36] ? En réalité, la question de la pérennité des supports physiques, des logiciels et des formats se pose sur le même mode en ce qui concerne les données et les méta-données.

D'autre part, la gestion des méta-données se heurte à trois écueils potentiels. Le premier est lié à ce que ces systèmes sont extensibles à l'infini. Les systèmes de méta-information sont en effet destinés à clarifier l'opacité des codifications formelles et à en réduire l'incertitude. A cette fin, la solution requise réside souvent dans l'utilisation d'une codification plus riche et donc, plus explicite : le langage naturel[37]. Cependant, le langage naturel est son propre métalangage. Dès lors, toute description émise sous cette forme peut faire l'objet d'un nombre infini de commentaires d'ordre supérieur. Ceci se traduit par la lourdeur et le coût de leur gestion pratique, lorsque celle-ci repose sur une mise à jour manuelle. Le second écueil tient à ce que les méta-données peuvent être elles-mêmes erronées et incertaines : leur validation ne peut faire l'objet systématique de test d'intégrité rigoureux. Le troisième écueil tient au décalage temporel entre la mise à jour d'une donnée et de la méta-donnée correspondante, cette dernière, surtout lorsqu'elle se présente sous une forme textuelle, n'étant généralement créée qu'au terme d'une phase d'analyse plus ou moins longue.

---

[32] http://ssdoo.gsfc.nasa.gov/nost/isoas/

[33] Masanes J., *Op. cit.* p. 14-18.

[34] De nombreuses applications « vivantes » incluent la gestion simultanée de vastes bases de données et des méta-données correspondantes, telles que, dans le domaine de la sécurité sociale, les glossaires incluant la documentation technique des déclarations trimestrielles DMFA et DRS. BOYDENS I., *E-gouvernement en Belgique : un retour riche d'expériences*. Techno Edition Spéciale, n°26, novembre 2003.

[35] *MoReq Specification, Model Requirement for the Management of Electronic Records,* IDA (Interchange Data between Administrations) Programme of the European Commission, mars 2001 (http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc, consulté le 18 mars 2004), p. 7.

[36] Walbel G., « Like Russian Dolls : Nesting Standards for Digital Preservation », *RLG DigiNews*, 15 juin 2003, vol. 7, n°3 (Web-based Newsletter : http://www.rlg.org/preserv/diginews/diginews7-3.html, consulté le 18 mars 2004).

[37] Langage parlé, évolutif et codifié par le seul usage.

A cet égard, les recommandations du « Web sémantique », développées sous l'égide du W3C, constituent une voie d'avenir: elles visent à décrire l'information grâce à des normes exploitables par une machine et compréhensibles par les humains[38]. Transversales à plusieurs domaines d'application, ces normes, dont la mise en place repose sur un consensus entre communautés d'utilisateurs, pourraient faciliter la création d'un réseau sémantique reliant données et méta-données[39].

Si les méta-données sont indispensables, il convient d'y recourir avec parcimonie et de privilégier les méta-données générées semi-automatiquement[40]. En outre, il convient d'agir parallèlement en « amont », lorsque l'on dispose d'une « prise » sur la source, lors de la constitution et de la gestion des données numériques, comme nous l'envisageons au point suivant.

4.6. Quelques principes de conception et de gestion

En un certain sens, le long terme prend naissance en t_0 dès la création des données. Quelques pratiques de conception et de gestion adoptées d'emblée permettent de favoriser la conservation des données :

- à propos des formats, choix de spécifications ouvertes et accessibles publiquement ;
- à propos des modalités de conception d'une application, recours au principe de la réutilisation, par exemple, via la mise en place du concept de « WOPM », « *Write Once Publish Many* »[41], qui consiste à générer une source unique (en XML) et à la diffuser sous divers formats (ASCII, PDF...). Ceci permet l'exploitation homogène des données à diverses fins (de façon à maintenir la cohérence entre les données transactionnelles et les méta-données correspondantes en ce qui concerne les codifications, par exemple) ;
- sur le plan de la gestion, il est conseillé :
  - d'éviter les couches de complexité qui rendent plus ardues les phases de migration (cryptage, compression), quand celles-ci ne sont pas indispensables ;
  - de déployer des stratégies d'amélioration de la qualité des données ; par exemple :
    - examen continu de la performance des processus ;
    - suivi de l'adéquation des données aux usages (ce qui inclut la détection de données « non utilisées », redondantes, etc.) ;
    - comparaisons croisées entre « *back ups* » ou entre fichiers concurrents répertoriant des données analogues (listes d'adresses, par exemple) issues de sources distinctes.

A propos du troisième point, en cas d'incohérence, une enquête quant à son origine doit être menée (altération d'un support, par exemple) et une nouvelle version « intègre » doit être générée. Dans d'autres cas, les incohérences témoignent d'une inadéquation conceptuelle entre le système d'information et le réel appréhendé (suite à une évolution de celui-ci) et la structure de l'application doit être revue.

---

[38] Boydens I., Du « Web sémantique » au « Web pragmatique », SmalS-MvM Research Note, mars 2004, n°5, p.19.

[39] Day M., « Integrating Metadata Schema Registries with Digital Preservation Systems to Support Interoperability : a Proposal » 2003 Dublin Core Conference, Seattle 28 septembre – 2 octobre 2003 (http://www.siderean.com/dc2003/Paper38-abstract.pdf), consulté le 29 février 2004.

[40] Par exemple, sur la base de listes pré-contrôlées de mots-clés spécifiant la signification des champs d'une base de données.

[41] Boydens I., « Déploiement coopératif d'un dictionnaire électronique de données administratives », *Revue Document Numérique* (« *Création et gestion coopératives de documents numériques d'information et de communication* »), vol. 5, nr 3-4/2001, p. 27-43.

5. Conclusions : de la pérennité au « continuum »

Les problèmes à la source de la détérioration de l'information numérique sont aussi variés qu'interdépendants[42]. La figure 2 présente un récapitulatif de ces interdépendances.

Parmi les difficultés à la source de la non conservation de l'information numérique, les questions liées à l'évolution des logiciels et des formats sont de loin les plus préoccupantes.

	hardware composantes	hardware périphériques	supports physiques	driver des supports	logiciels logiciel applicatif	logiciels Operating system	formats
hardware (composantes)		X	X	X	X	X	
hardware (périphériques)	X		X	X	X	X	
supports physiques	X	X		X	X	X	
driver des supports	X	X	X			X	
logiciels (applicatifs)	X	X	X			X	X
logiciel (Operating System)	X	X	X	X	X		
formats					X		

Figure 2. Récapitulatif des interactions entre les facteurs à l'origine d'une détérioration de l'information numérique.

La figure 3 synthétise la correspondance entre les stratégies de préservation et les problèmes qu'elles sont censées résoudre. Si la mise en place de méta-données est fondamentale, ces dernières sont elles-mêmes des données et sont dès lors tout aussi fragiles.

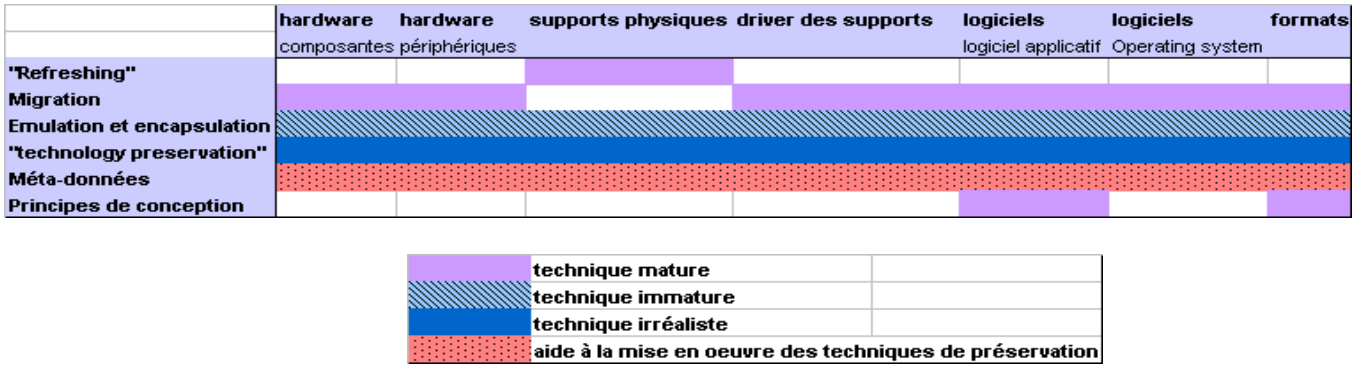

Figure 3. Les techniques de préservation de l'information numérique.

Nous avions défini au seuil de l'article trois « moments » dans le cycle de vie des données : le « temps court » des données de gestion, le « temps intermédiaire » de la préservation légale et le « temps long » de la conservation historique. La figure 4 ci-dessous synthétise le recours préconisé aux différentes stratégies en fonction de l'état d'une donnée dans son cycle de vie.

[42] Picart K., De la conservation à long terme de l'information numérique, mémoire de fin d'études, Université Libre de Bruxelles, 2000.

	temps court (2 ans)	temps intermédiaire (5 à 10 ans)	temps long (10 ans et plus)
"Refreshing"	X	X	X
Migration	X	X	X
Emulation et encapsulation			si maturité
"technology preservation"			
Méta-données	X	X	X
Principes de conception	X		

Figure 4. les stratégies à mettre en oeuvre en fonction du cycle de vie des données.

Une combinaison du « *refreshing* » et de la migration s'impose tout au long du cycle de vie des données.

S'agissant de la gestion de données « vivantes » se déployant « dans le temps court », des stratégies complémentaires contribuent à la conservation de l'information, telles que par exemple, la mise en place d'applications dont la structure est économe (selon le principe du WOPM, notamment) ou le déploiement de stratégies en vue de contrôler la qualité des données. Ces démarches de qualité requièrent une organisation structurée : elles visent non seulement à préserver l'intégrité de l'information mais aussi à en assurer la conservation à long terme. Il s'agit également de choisir des normes ouvertes et documentées pour les formats (telle que la norme XML du W3C) et le codage.

Plus l'environnement d'une donnée est simple, plus celle-ci est « pérenne » dans la mesure où il y a moins de couches de complexité à traiter pour la conservation à long terme. C'est ainsi, par exemple, que les microfilms sont parfois envisagés pour le long terme en dépit de leur « passivité fonctionnelle ». Cela dit, en ce qui concerne les larges volumes de données structurées, un medium numérique reste requis.

Dans tous les cas, vu l'ampleur de la problématique, il est préférable, s'agissant de l'information, de parler de « *continuum* » plutôt que de « pérennité ». En effet, il semble illusoire de figer à tout jamais les données « en l'état » sans altération aucune. Des pertes de données peuvent par exemple survenir au cours des opérations de migration, pourtant indispensables. On observe dès lors une forme de « *continuum* » : les données conservées évoluent d'une façon ou d'une autre dans le temps.